Understanding Bayesian Credible Intervals

Mike Dehn

10/13/2021

1 Introduction

In this tutorial, we will explore the idea of Bayesian credible intervals. For complete context, we begin with an introduction to both the frequentist and Bayesian approaches to statistical thinking. We further consider the concept of confidence and credible intervals in each approach. Once this conceptual foundation is established, we will walk through an example of applying these concepts to real world data.

1.1 Frequentist vs. Bayesian Statistical Inference

Assume that there is a parameter, ϕ, that we would like to estimate for a population. As an example, you might consider the average age of the people living in the United States. While the United States collects thorough census data on a recurring basis, it is very difficult to know whether the data are complete. Some people may not respond to the census, others may be dishonest in their responses. The key is, we cannot know the value of ϕ with absolute certainty; we must estimate based on the data available to us. Statistical inference is a process that allows us to build these estimates based on the data we have about a given population.

There are two primary approaches to statistical inference: frequentist and Bayesian. In frequentist statistics, samples from a population are used to estimate the value of the parameter in question, but we do not assume that parameter is a random variable. Instead, we assume that the parameter has a fixed but unknown value that is not stochastic in nature. A frequentist might say, "the probability that the average age of people in the United States is 30 years old is either 0% or 100%, I just don't know which one is correct."

Conversely, in Bayesian statistics, the parameter in question is assumed to be a random variable. A Bayesian statistician might say, "the probability that the average age of people in the United States is 30 years old is X%, based on what we know from population data that we have on hand." The Bayesian statistician can provide a more precise estimate of the probability of a particular value for ϕ, while the frequentist cannot. Note that this does not necessarily imply that the Bayesian is more accurate!

NOTE: If further explanation is helpful, there is an excellent article and video on the Towards Data Science website that provides an intuitive explanation of the differences between Bayesian and frequentist statistics: https://towardsdatascience.com/statistics-are-you-bayesian-or-frequentist-4943f953f21b

Both forms of statistical inference are rife with uncertainty. We do not *know* the true value of ϕ based on our statistical procedures; what we *do have* is an estimate with an established level of uncertainty. In frequentist statistics, this certainty is described using confidence intervals. Bayesian statistics uses the related, but different, credible interval.

1.1.1 Frequentist Confidence Intervals

The frequentist notion of confidence intervals is generally well-known in most communities with some degree of statistical sophistication, but are often misunderstood. In particular, the meaning of the confidence level

is perhaps the most confusing. Many people incorrectly believe that a confidence interval is a range of values that contain the estimated parameter with a probability equal to the confidence level. In other words, if the 95% confidence interval for ϕ is such that 25 <= ϕ <=35, it is tempting to assert that there is a 95% probability that ϕ lies between 25-35. You may be asking, "why isn't this true?"

As explained in the previous section of this tutorial, frequentist statistics does not assignprobabilities to parameter estimates - frequentists assume that the *true* value of the parameter is not a random variable, rather it is an unknown non-stochastic value.

What, then, does a confidence level signify? The definition is not straightforward: "a confidence level represents the theoretical long-run frequency (i.e., the proportion) of confidence intervals that contain the true value of the unknown population parameter."

This is a little confusing, so let's break it down. In practical terms, this definition says: Let's imagine we can run the same experiment over and over again. If you do, in the long-run, you'll find that the percentage of confidence intervals that contain ϕ tends toward the confidence level percentage we selected. For a 90% confidence level, 90% of confidence intervals derived from repeats of this experiement will contain the true value of ϕ.

If this doesn't feel super satisfying to you, you aren't alone. The frequentist confidence interval sidesteps assigning a probability to the true value of ϕ in favor of making an assertion with the confidence we can have as a result of the experiment itself.

1.1.2 Bayesian Credible Intervals

So, is there a way to establish the more intuitive notion of a probability estimate for the value of ϕ itself within an interval? If you are a Bayesian statistician, the answer is "yes."

Bayesian inference uses the notion of *credible intervals* to assign a probability that the estimated parameter ϕ lies within an established range. When we calculate the 95% credible interval for a sample, we assert that the probability of the true value of ϕ resides in the given interval is 95%, given the data available to us.

This definition speaks to many peoples' intuition about what an estimate interval *ought* to be. It maps well to our understanding of probability as an estimation of what we think is true, while providing an objective accounting for our uncertainty in the estimate.

Although credible intervals are more intuitively understood than their frequentist counterparts, they do have some undesirable characteristics. For example, they require us to make assumptions about a prior distribution. They also require us to construct posterior distribution, which can be computationally expensive.

2 Learning by Example with Real World Data

With this foundational introduction, let us continue with a concrete example of calculating a credible interval for a data set in R. In this example, we will use a data set collected by the United States Center for Disease Control (CDC) as part of their National Health and Nutrition Examination Survey for calendar years 2017-2020. In particular, we will establish a credible interval for the average BMI of adult males in the United States.

One advantage to using this data set and estimating the mean BMI parameter is that it is a well-studied parameter within the healthcare community. Therefore, we can establish some intuition as to the accuracy of our result.

2.1 Ingesting Data

The data set can be downloaded from the CDC's website. There are two files that we need to download:

- Body measures - https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_BMX.XPT
- Demographics - https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.XPT

The body measures data contains height, weight, BMI, and related metrics for each participant in the survey. The demographics data provides various biographical data for each participant, which we will use to filter only for adult males. According to the CDC's website, the data set considers "adults" to be age 20 or higher at the time of their measurements.

We can programmatically download the data set from the CDC's website using the code below. If this snippet does not work for you, try downloading the data manually.

```
curl::curl_download('https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_BMX.XPT',
                    'body_measures.xpt')

curl::curl_download('https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.XPT',
                    'demographics.xpt')
```

You should now have two files in your working directory - `body_measures.xpt` and `demographics.xpt`. The data are in a non-human readable data format, but they can be ingested using the `haven` R package.

```
library(tidyverse)
library(haven)
body_measures <- read_xpt('body_measures.xpt')
demographics <- read_xpt('demographics.xpt')
```

Now that we have ingested the body measures data set, we should check that the results make sense. Let's take a look at a summary of the data set.

```
summary(body_measures)
```

```
##       SEQN           BMDSTATS         BMXWT            BMIWT
##  Min.   :109263   Min.   :1.000   Min.   :  3.20   Min.   :1.000
##  1st Qu.:113171   1st Qu.:1.000   1st Qu.: 42.30   1st Qu.:3.000
##  Median :117082   Median :1.000   Median : 68.10   Median :3.000
##  Mean   :117069   Mean   :1.135   Mean   : 65.43   Mean   :2.993
##  3rd Qu.:120974   3rd Qu.:1.000   3rd Qu.: 86.30   3rd Qu.:3.000
##  Max.   :124822   Max.   :4.000   Max.   :254.30   Max.   :4.000
##                                   NA's   :225      NA's   :13712
##     BMXRECUM        BMIRECUM        BMXHEAD         BMIHEAD
##  Min.   : 49.10   Min.   :1       Min.   :32.40   Min.   : NA
##  1st Qu.: 70.10   1st Qu.:1       1st Qu.:39.20   1st Qu.: NA
##  Median : 82.95   Median :1       Median :41.30   Median : NA
##  Mean   : 81.51   Mean   :1       Mean   :41.12   Mean   :NaN
##  3rd Qu.: 92.88   3rd Qu.:1       3rd Qu.:42.98   3rd Qu.: NA
##  Max.   :113.90   Max.   :1       Max.   :48.30   Max.   : NA
##  NA's   :12830    NA's   :14257   NA's   :13990   NA's   :14300
##      BMXHT           BMIHT           BMXBMI         BMDBMIC         BMXLEG
##  Min.   : 78.3   Min.   :1.000   Min.   :11.90   Min.   :1.00   Min.   :24.80
##  1st Qu.:151.1   1st Qu.:1.000   1st Qu.:20.40   1st Qu.:2.00   1st Qu.:36.00
##  Median :162.1   Median :3.000   Median :25.80   Median :2.00   Median :38.90
##  Mean   :156.5   Mean   :2.181   Mean   :26.66   Mean   :2.55   Mean   :38.75
##  3rd Qu.:171.3   3rd Qu.:3.000   3rd Qu.:31.40   3rd Qu.:3.00   3rd Qu.:41.50
##  Max.   :199.6   Max.   :3.000   Max.   :92.30   Max.   :4.00   Max.   :55.00
```

```
##  NA's   :1143    NA's   :14129   NA's   :1163    NA's   :9551   NA's   :3316
##      BMILEG          BMXARML         BMIARML          BMXARMC        BMIARMC
##  Min.   :1       Min.   : 9.4   Min.   :1       Min.   :11.20   Min.   :1
##  1st Qu.:1       1st Qu.:31.9   1st Qu.:1       1st Qu.:23.60   1st Qu.:1
##  Median :1       Median :36.0   Median :1       Median :30.30   Median :1
##  Mean   :1       Mean   :33.8   Mean   :1       Mean   :29.32   Mean   :1
##  3rd Qu.:1       3rd Qu.:38.6   3rd Qu.:1       3rd Qu.:34.90   3rd Qu.:1
##  Max.   :1       Max.   :49.9   Max.   :1       Max.   :64.50   Max.   :1
##  NA's   :13812   NA's   :810    NA's   :13813   NA's   :816     NA's   :13807
##     BMXWAIST         BMIWAIST         BMXHIP          BMIHIP
##  Min.   : 40.00   Min.   :1       Min.   : 62.5   Min.   :1
##  1st Qu.: 73.30   1st Qu.:1       1st Qu.: 95.5   1st Qu.:1
##  Median : 91.00   Median :1       Median :103.4   Median :1
##  Mean   : 89.67   Mean   :1       Mean   :105.7   Mean   :1
##  3rd Qu.:105.40   3rd Qu.:1       3rd Qu.:113.3   3rd Qu.:1
##  Max.   :187.50   Max.   :1       Max.   :187.5   Max.   :1
##  NA's   :1726     NA's   :13683   NA's   :4438    NA's   :13924
```

The most important variable in this data set is the BMXBMI variable, which corresponds to the BMI calculated for each individual under observation in the survey. We can see that there are 1,163 rows containing `NA` values for this variable. In our calculations, we will simply reject any rows containing `NA` for the BMXBMI variable, but it is important for us to remember this in the context of describing any uncertainty in our final result.

The code snippet below will perform this cleaning for us.

```
library(dplyr)
body_measures_clean <- filter(body_measures, ! is.na(BMXBMI))
```

We no longer have any `NA` values in the BMI field, so we have rejected any samples that may be problematic for our parameter estimation. However, we have not yet selected for the specific population of interest: 20+ year old males.

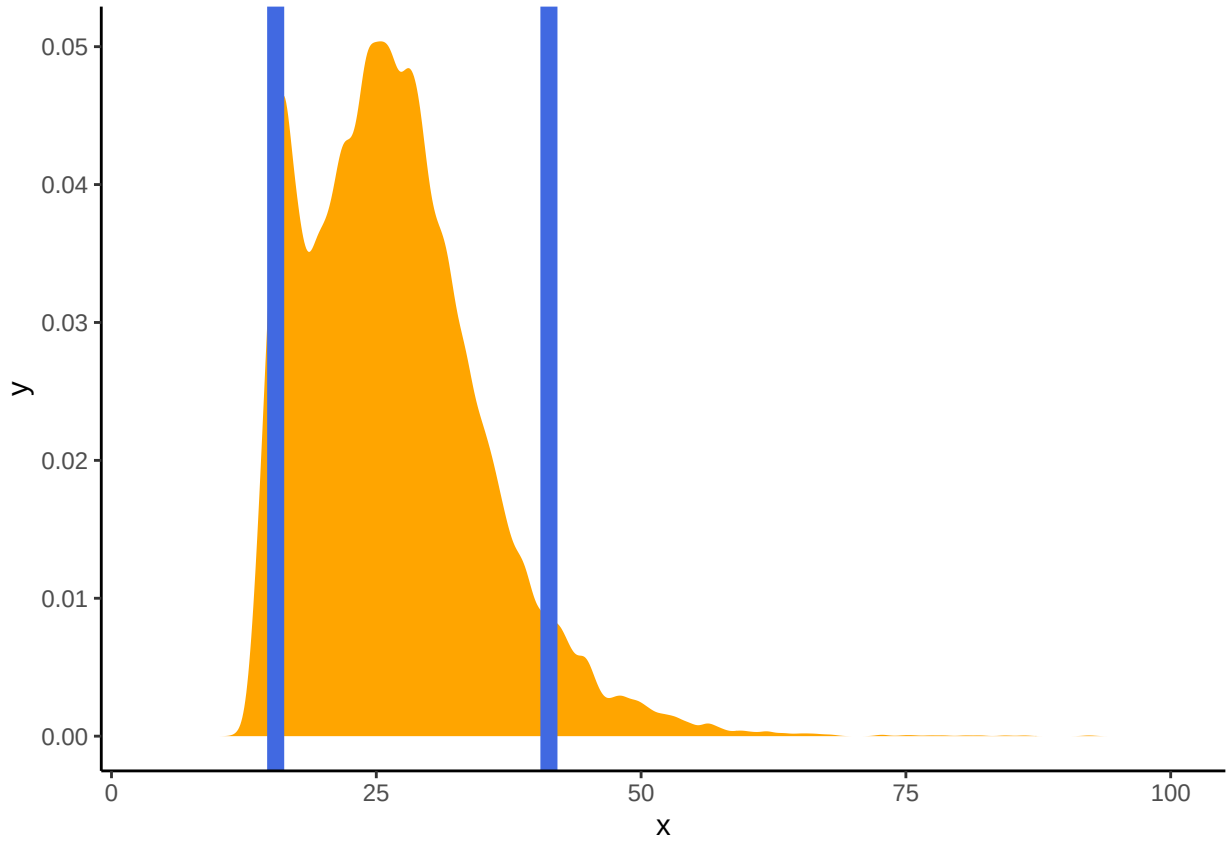
To select on this particular subset of the population, we can join the `body_measures` and `demographics` data set on their common column `SEQN` that uniquely identifies a participant. We then filter any rows do not satisfy the criteria of being males of age 20 or greater.

```
adult_male_data <- inner_join(body_measures_clean, demographics, by = 'SEQN')
adult_male_data <- filter(adult_male_data, RIAGENDR == 1, RIDAGEYR >= 20)
```

2.2 Calculating Credible Interval

```
library(bayestestR)
ci_hdi <- ci(body_measures_clean$BMXBMI, method = "ETI", ci = 0.89)

body_measures_clean$BMXBMI %>%
  estimate_density(extend=TRUE) %>%
  ggplot(aes(x = x, y = y)) +
  geom_area(fill = "orange") +
  theme_classic() +
  # HDI in blue
  geom_vline(xintercept = ci_hdi$CI_low, color = "royalblue", size = 3) +
  geom_vline(xintercept = ci_hdi$CI_high, color = "royalblue", size = 3)# +
```




```
# Quantile in red
#geom_vline(xintercept = ci_eti$CI_low, color = "red", size = 1) +
#geom_vline(xintercept = ci_eti$CI_high, color = "red", size = 1)
```

2.3 Plotting Credible Interval

2.4 Interpreting the Credible Interval and Reporting Uncertainty

3 Conclusions